# DK CLARIN WP3.2 - konverteringsprogrammer

Jensen, Torben Juel

*Publication date:*
2010

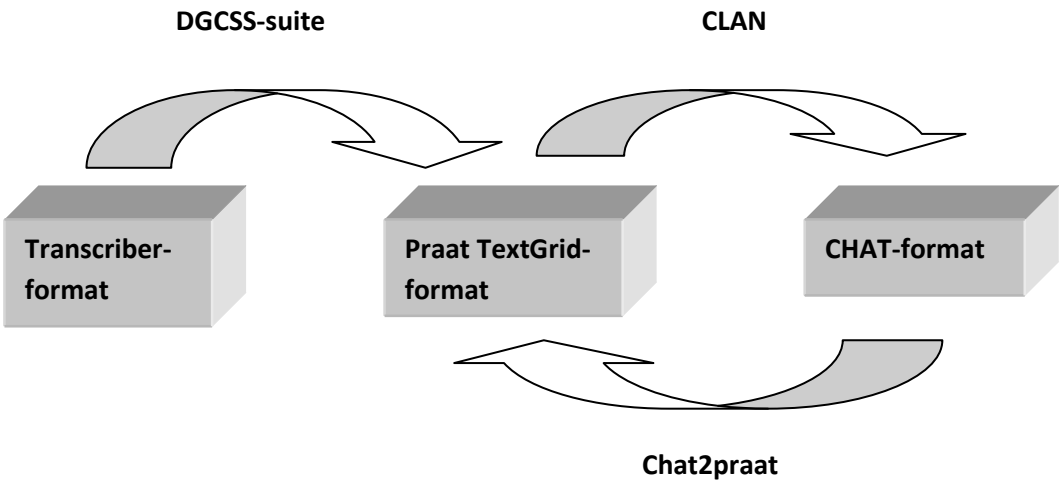
*Document version*
Også kaldet Forlagets PDF

# Oversigt over WP3.2-konverteringsprogrammer

En del af målet for WP3.2. "Tool box for the treatment of spoken materials for various types of research" er at etablere en pakke af konverteringsprogrammer mellem de mest anvendte formater til transskription og annotering af talt sprog.

LANCHART har hertil udviklet to nye programfaciliteter, DGCSS-suite (skrevet af Peter Lind) og Chat2praat (skrevet af Jakob M. Christensen), og har indgået i et samarbejde med Brian MacWhinney og Leonid Spektor, Carnegie Mellon University, der har ført til udvikling af en ny konverteringsfunktion i det eksisterende program CLAN (Computerized Language ANalysis).

Tilsammen gør de tre programmer det muligt at konvertere mellem formaterne *Transcriber*, *Praat TextGrid* og *CHAT*:




DGCSS-suite kan desuden konvertere filer i *KUAtekst-format* samt det format Peter Juel Henrichsen har udviklet netversionen af BySoc-korpusset, til CHAT-format (Lind & Jensen 2006).

Endelig har LANCHART udviklet et program (skrevet af Jakob M. Christensen) til realignment af lydfil og annotering i Praat-TextGrid der er genereret som resultat af konvertering af filer i Transcriber- eller CHAT-format udskrevet uden alignment af lyd og transskription.

# DGCSS-suite

DGCSS-suite er et java-program og er distribueret som en .jar-fil. Det betyder at enhver computer, uanset styresystem, som har en Java Virtual Machine, vil kunne afvikle programmet.

DGCSS-suite har tre inputformater:

1) Partiturformatet til linjeeditoren KUATEKST, der blev udviklet i forbindelse med SHF-initiativet "Dansk talesprog i dets variationer", og som er benyttet blandt andet i Projekt Bysocioling-vistik (Gregersen, Albris & Pedersen 1991) og Tore Kristiansens Næstved-undersøgelser (Kristiansen 1998).

2) Det specialformat der danner grundlag for netversionen af BySoc-korpusset (http://bysoc.dyndns.org/), som er udviklet af Peter Juel Henrichsen. Dette format er blandt andet kendetegnet ved at transskriptionen af hver talers tale lagres som separate filer (Henrichsen 1997; Allwood m.fl. 2005).

3) Transcriber (http://trans.sourceforge.net/en/presentation.php).

I DGCSS-suite konverteres inputformatet indledningsvist uden informationstab til et XML-super-format. Superformatet eksisterer fortrinsvist internt i computerens hukommelse under konverte-ringen, men det er muligt at gemme det i form af en XML-fil. I XML-superformatet er det i DGCSS-suites editor muligt at tilføje eller ændre header-oplysninger og deltagerkoder inden transskriptio-nen konverteres til det valgte output-format (Lind & Jensen 2006).

DGCSS-suite har to output-formater: CHAT (http://childes.psy.cmu.edu/) og Praat TextGrid (http://www.fon.hum.uva.nl/praat/).

DGCSS-suite kan dog p.t. kun i forbindelse med input-formatet Transcriber konvertere til Praat TextGrid-format. Konverteringen til TextGrid er kun kvalitetskontrolleret i forhold til optagelser udskrevet efter DGCSS' Transcriber-udskrivningskonvention (Jensen m.fl. 2009).

# Chat2praat

DGCSS-suite er et java-program og er distribueret som en .jar-fil. Det betyder at enhver computer, uanset styresystem, som har en Java Virtual Machine, vil kunne afvikle programmet.

Chat2praat har CHAT som input-format. Dette format konverteres uden informationstab bortset fra header-informationer til et Praat TextGrid (TextGrid har ingen header), dog således at formelle fejl og tvetydigheder i forbindelse markeringer af overlappende tale i CHAT-filen må ændres først ud fra Chat2praats fejlmeddelelser.

Chat2praats konverteringsproces er kun kvalitetskontrolleret i forhold til optagelser udskrevet efter DGCSS' CHAT-udskrivningskonvention (Pharao m.fl. 2006).

# CLAN (PRAAT2CHAT)

CLAN er udviklet af Leonid Spektor og Brian MacWhinney til at analysere data udskrevet i CHILDES(Child Language Data Exchange System)-formatet CHAT (MacWhinney 2009; MacWhinney 2010).

LANCHART har samarbejdet med Spektor og MacWhinney (som medvirker i WP3) om udviklingen af en funktion i CLAN, "PRAAT2CHAT", der gør det muligt at konvertere Praat TextGrid til CHAT-format (MacWhinney 2010). Udviklingen af funktionen har taget udgangspunkt i DGCSS' specifikke udnyttelse af Praats TextGrid-format (jf. bilag 1), og LANCHART har kun kvalitetskontrolleret konverteringen i forhold til DGCSS' annoteringskonventioner.

Konverteringsprocessen fra Praat TextGrid til CHAT sker ikke helt uden informationstab pga. uforeneligheder mellem CHAT- og TextGrid-formatet. Disse skyldes primært at annotering i Praat TextGrid alene defineres ved tidskoder (som intervaller med xmin- og xmax-værdier), mens annotering i CHAT er defineret i relation til det hovedtier (*the main line*) som indeholder den (ytringsopdelte) transskription af optagelsen. Dette kan give problemer i forbindelse med repræsentationen i CHAT af visse typer af overlappende tale (dvs. markeringen af hvilke dele af to eller flere ytringer der overlapper tidsligt med hinanden). Desuden er markeringen af start- og slutpunkter i CHAT-transskriptionen af annoteringer der ikke knytter sig specifikt til en enkelt talers tale (og som i CHAT markeres vha. *gems*), ikke i alle tilfælde helt præcis i forhold til det TextGrid der danner udgangspunkt for konverteringen.

# Bilag 1: Description of the DGCSS Praat format

The Praat TextGrids containing transcription and annotation of the conversations are comprised of a number of tiers of which the "ortografi"-tiers are basic: Each word (whether completed or non-completed) identified in the recording is transcribed in the label of a unique interval in the "ortografi"-tier of the particular speaker. A closed set of *significant* vocal sounds, e.g. filled pauses, are also transcribed in the labels of the "ortografi"-tiers (see the list below).

Each speaker participating in the conversation is assigned a unique "ortografi"-tier named "ortografi (XXX)", of which the X's in the parenthesis constitutes a unique informant code comprised of three capital letters. Overlaps between two or more speakers are represented by alignment of the relevant intervals in the "ortografi"-tiers (and all dependent tiers) of the speakers in question.

All other tiers in the TextGrid are dependent on the "ortografi"-tiers: every boundary is aligned with a boundary in one or more of the "ortografi"-tiers, i.e. their intervals are aligned with one or more intervals in at least one "ortografi"-tier. Tiers specifically connected to the "ortografi"-tier of a particular speaker include the speaker's informant code in their names in the same way as "ortografi"-tiers, e.g. "events (XXX)". Tiers containing information about the conversation as such - in a particular passage - are named without informant codes, e.g. "Samtaletype".

Every speaker in a conversation is assigned the following 6 tiers: "ortografi", "phonetic", "emphasis", "events", "uncertain transcription" and "comments". Thus, every TextGrid contains these tiers – multiplied by the number of speakers participating in the particular conversation. The content of the tiers is described below:

**"ortografi"-tier:**

| **Phenomenon** | **Representation (in the label of a separate interval)** |
|---|---|
| Completed word | Danish standard orthography<br><br>- with the exception that hyphen (in acronyms and complex words) and slash is replaced by underscore |
| Uncompleted word | hyphen: "xxx-" |
| Resumption of uncompleted word | hyphen: "-xxx" |
| Unintelligible passage | Three (lower-case) x's: "xxx" |
| Laughter | "ha" |
| Significant in- or exhalation | "hh" |
| Affirmation/backchanneling | "mm", "nja" or "njo" |
| Pause | empty |
| Filled pause | "øh" |
| Shush | "sh" |
| Exclamation | "uh" |

**"phonetic"-tier (intervals aligned with *one* interval in an "ortografi"-tier):**

| **Phenomenon** | **Representation (in the label of a separate interval)** |
|---|---|
| lengthening of vowel or consonant | ":" immediately after the letter representing the lengthened sound (the word transcribed in the "ortografi"-tier is repeated with a colon): "ja:", "m:åske" |

**"emphasis"-tier (intervals aligned with *one* interval in an "ortografi"-tier) :**

| **Phenomenon** | **Representation (in the label of a separate interval)** |
|---|---|
| emphasis | "!" |

**"events"-tier (intervals aligned with *one* interval in an "ortografi"-tier):**

| **Phenomenon** | **Representation (in the label of a separate interval)** |
|---|---|
| sound – non-word which cannot be described as one of the conventionalized meaning bearing sounds defined above | description of sound (aligned with an *empty* interval in the ortografi-tier): "host", "støn" |

**"uncertain transcription" (intervals aligned with one or more intervals in an "ortografi"-tier)**

| **Phenomenon** | **Representation (in the label of a separate interval)** |
|---|---|
| uncertain transcription | "?" |

**"comments" (intervals aligned with one or more intervals in an "ortografi"-tier, can be global)**

| **Phenomenon** | **Representation (in the label of a separate interval)** |
|---|---|
| transcribers comment | e.g. "hvisker det sidste", "snakker til hunden" |

**Tiers containing *analytical* information**

In the *analytical* annotation processes a number of tiers are added to the TextGrids. The list of these tiers is still open and the number of tiers varies from TextGrid to TextGrid, depending on how far it has progressed in the analytical processes. These tiers are all dependent on the "ortografi"-tiers, and they can be divided into three sub-groups:

1. *Global* tiers (without informant codes): Tiers containing information about the conversation as such - in a particular passage. They are not connected to a particular speaker, and the first boundary of

an interval may be aligned with a boundary in one speaker's "ortografi"-tier, whereas the last boundary may be aligned with a boundary in another speakers "ortografi"-tier. E.g. "Inter-aktionsstruktur" and "Samtaletype".

2. Speaker specific tiers with intervals aligned to *a sequence of intervals* in the relevant "ortografi"-tier, e.g. "ledsaet (XXX)"
3. Speaker specific tiers with intervals aligned to *a single interval* in the relevant "ortografi"-tier, i.e. one-to-one with a word. E.g. "parole_PoS (XXX)" and "Grammatik (XXX)".

# Referencer

Allwood, Jens; Peter Juel Henrichsen; Leif Grönqvist; Elisabeth Ahlsén & Magnus Gunnarsson (2005). "Transliteration between spoken language corpora." Nordic Journal of Linguistics **28**(1): 5-36.

Gregersen, Frans; Jon Albris & Inge Lise Pedersen (1991). Data and design of the copenhagen study. The copenhagen study in urban sociolinguistics. Frans Gregersen & Inge Lise Pedersen. København, C. A. Reitzels Forlag.

Henrichsen, Peter Juel (1997). "Talesprog med ansigtsløftning." Instrumentalis **10**.

Jensen, Torben Juel; Janus Møller; Line Visby Nielsen; Nina Kanstrup Hansen & Louise Gad (2009). Transcriber - dgcss' udskrivningsmanual, Danmarks Grundforskningsfonds center for sociolingvistiske sprogforandringsstudier.

Kristiansen, Tore (1998). "The role of standard ideology in the disapperance of the traditional danish dialects." Folia Linguistica **XXXII(1-2)**: 115-129.

Lind, Peter & Torben Juel Jensen (2006). Konvertering af filer fra bysoc til dgcss-chat, Danmarks Grundforskningsfonds center for sociolingvistiske sprogforandringsstudier.

MacWhinney, Brian (2009). The childes project. Tools for analyzing talk – electronic edition. **2, part 1**.

MacWhinney, Brian (2010). The childes project. Tools for analyzing talk – electronic edition. **2, part 2**.

Pharao, Nicolai; Torben Juel Jensen; Minna Olesen; Malene Monka; Janus Møller; Inge Lise Pedersen; Jens Norman Jørgensen & Frans Gregersen (2006). Dgcss' udskrivnings- og korrekturmanual (clan), Danmarks Grundforskningsfonds center for sociolingvistiske sprogforandringsstudier.